Extraction of medical information from textual sources: a statistical variant of the Boundary Word method

Terrell W. Herzig, M.S.H.I., Merida Johns, Ph.D.
University of Alabama at Birmingham
Birmingham, Alabama

Background. The basis of this project began six months ago from related work on a conceptual analysis of the feasibility of extracting medical information from electronic patient charts and translating it into expert systems such as DXPLAIN.

Early analysis focused on the ability to automatically extract relevant medical information from patient charts, nursing notes, pre-admission exams, summaries, etc. Once this was accomplished, the information would be compared against the vocabularies in the UMLS knowledge sources for recording and coding discrepancies. With additional algorithms and the utilization of the UMLS semantic network, the information could be re-targeted to a new vocabulary. In theory, information could then be collected and processed by expert systems without intervention. To accomplish our goals, it would be necessary to develop algorithms capable of processing natural language.

Traditional natural language processing often utilizes a technique, known as conceptual graphs, to break down a text and identify concepts for further processing. This technique will "parse" a segment of text, extracting words and building what is known as a conceptual graph. As a conceptual graph, meaning of the generated tokens and their position within the graph help programs infer semantic relationships. The conceptual graph, can then be used to extract meaning from the sentence, when compared against a programs knowledge base.

One serious difficulty to this approach, when applied to medicine, is the inability of this technique to recognize new vocabulary terms at first occurrence. New terms are frequently being added to medical vocabularies and lexicons. This is evidenced by the growth within the UMLS Meta-thesaurus from its original inception and its updating on an annual basis.

These shortfalls, coupled with the maintenance and usage of large vocabularies, was a distraction to implementing an algorithm for the extraction of patient information based on conceptual graph theory. Our search for alternative techniques led us to a technique developed in 1988 called the "Boundary Word Method". This technique relied on words with low biomedical informational content (designated as "barrier words") separating important multiple-word biomedical information. Continued research uncovered a revival of the technique in 1995 for purposes of recognizing concepts in medical texts with a fine level of granularity.

System. Our research focused on the development of a parser program utilizing a statistical variation of the "Boundary Word Method." The parser was constructed in a modular fashion to allow replacement of source streams, symbol tables, boundary word lists, output streams, and statistical modules at run time. Statistical programs were incorporated to allow the parser to "determine" appropriate content without the use of long word lists, inherit to the boundary word technique.

Evaluation. Twenty five examples each of discharge summaries, nursing notes, pre and post exams were collected in electronic format from transcribed medical records. Each document was at least 5 pages in length. Each file was parsed and evaluated to determine candidate phrases and structural granularity.

Results. We expected the technique to produce lists with similar granularity to the traditional technique but with a smaller word list (derived statistically) and requiring less processing time. Initial results indicate authentic word candidacy approaching 77% with the same fine granularity as the 1995 technique. Word list size was reduced from thousands of boundary words to a few hundred.

1091—8280/97/$5.00